

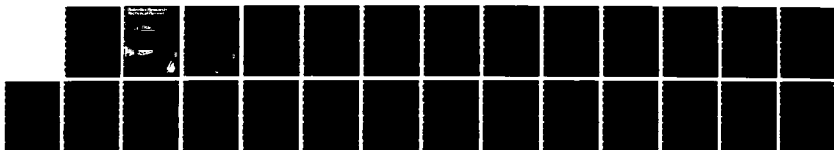
AD-A178 958

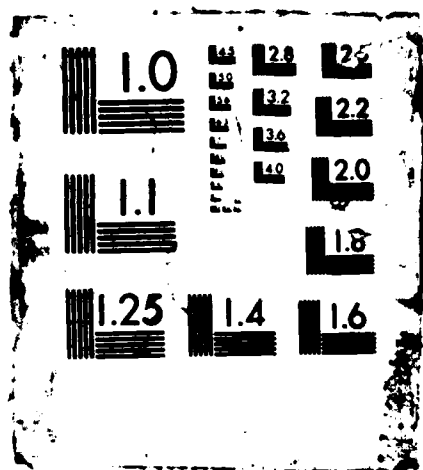
A STATISTICAL VIEWPOINT ON THE THEORY OF EVIDENCE
REVISION(U) NEW YORKUNIV NY COURANT INST OF
MATHEMATICAL SCIENCES R A HUMMEL ET AL MAY 86 TR-194
UNCLASSIFIED N00014-85-K-0077

1/1

F/G 9/4

NL





Robotics Research Technical Report

②

DTIC FILE COPY

Generatum omnis laboris ex machina

AD-A178 958

A Statistical Viewpoint on the Theory of Evidence

by

Robert A. Hummel †
Courant Institute of Mathematical Sciences

Michael S. Landy ‡
Department of Psychology

Technical Report No. 194
Robotics Report No. 57
December, 1985
Revised: May, 1986

This document has been approved
for public release and distribution
in unlimited quantities.

DTIC
ELECTE
APR 10 1987
S D E

New York University
Courant Institute of Mathematical Sciences

Computer Science Division
251 Mercer Street New York, N.Y. 10012

87-46083

Accession For	
NTIS GRA&I	<input checked="" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	<i>per</i>
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	



A Statistical Viewpoint on the Theory of Evidence

by

Robert A. Hummel †
Courant Institute of Mathematical Sciences

Michael S. Landy ‡
Department of Psychology

Technical Report No. 194
Robotics Report No. 57
December, 1985
Revised: May, 1986

†New York University
251 Mercer Street
New York, New York 10012
hummel@nyu.arpa

‡New York University
6 Washington Place
New York, New York 10003
landy@nyu-acf8.arpa

DTIC
ELECTE
S APR 10 1987 D
E

This research was supported by Office of Naval Research Grant N00014-85-K-0077 and NSF Grant DCR-8403300. We thank Olivier Faugeras for an introduction to the topic, and Tod Levitt for much assistance. We appreciate the helpful comments given by George Reynolds, Deborah Strahman, and Jean-Claude Falmagne. Document preparation was done by Linda Narcowich.

This document has been approved
for public release and sale; its
distribution is unlimited.

A Statistical Viewpoint on the Theory of Evidence

Robert Hummel
Michael Landy

Abstract

We describe a viewpoint on the Dempster/Shافر "Theory of Evidence", and provide an interpretation which regards the combination formulas as statistics of the opinions of "experts". This is done by introducing spaces with binary operations that are simpler to interpret or simpler to implement than the standard combination formula, and showing that these spaces can be mapped homomorphically onto the Dempster/Shافر theory of evidence space. The experts in the space of "opinions of experts" combine information in a Bayesian fashion. We present alternative spaces for the combination of evidence suggested by this viewpoint.

1. Introduction

Many problems in artificial intelligence call for assessments of degrees of belief in propositions based on evidence gathered from disparate sources. It is often claimed that probabilistic analysis of propositions is at variance with intuitive notions of belief [1,2,3]. Various methods have been introduced to reconcile the discrepancies, but no single technique has settled the issue on both theoretical and pragmatic grounds.

One method for attempting to modify probabilistic analysis of propositions is the Dempster/Shافر "Theory of Evidence." This theory is derived from notions of upper and lower probabilities, as developed by Dempster in [4]. The idea that intervals instead of probability values can be used to model degrees of belief had been suggested and investigated by earlier researchers [5,6,2,7], but Dempster's work defines the upper and lower points of the intervals in terms of statistics on set-valued functions defined over a measure space. The result is a collection of intervals defined for subsets of a fixed labeling set, and a combination formula for combining collections of intervals.

Dempster explained in greater detail how these notions could be used to assess beliefs on propositions in [8]. The topic was taken up by Shafer [9,10], and led to publication of a monograph on the "Theory of Evidence," [11]. All of these works after [8] emphasize the values assigned to subsets of propositions (the "beliefs"), and the combination formulas, and de-emphasize the connection to the statistical foundations based on the set-valued functions on a measure space. This paper will relate the statistical foundations of the Dempster/Shافر theory of evidence to notions of beliefs on propositions.

The Dempster/Shافر theory of evidence has sparked considerable debate among statisticians and "knowledge engineers". The theory has been criticized and

debated in terms of its behavior and applicability, e.g. [12, 13, 8 (Commentaries following)]. Some of the questions have been answered by Shafer [14, 15], but discussion of the theoretical underpinnings continues, e.g. [1, 16, 3].

Recently, there has been increased interest in the use of the Dempster/Shافر theory of evidence in expert systems [17, 18]. Most of the recent attempts to map the theory to real applications and practical methods, such as described in [19, 20, 21, 22, 23], are based on the techniques described by Shafer [11], and disregard the statistical theoretical foundations from which the theory was derived. In this paper we present a viewpoint on the Dempster/Shافر theory of evidence that regards the theory as statistics of opinions of "experts". We relate the evidence-combination formulas to statistics of experts who perform Bayesian updating in pairs. Finally, we suggest a related formulation that leads to simpler formulas and fewer variables.

Recently, the authors have pointed out that the Dempster/Shافر theory of evidence is one technique in a large class of iterative knowledge aggregation methods [24]. These methods, which include relaxation labeling [25], stochastic relaxation [26] and neural models [27], always attempt to find a true labeling by updating a state as evidence is accumulated. In the theory of evidence, as in many other models, the true labeling is one of a finite number of possibilities, but the state is a collection of numbers describing an element in a continuous domain. In the Shafer formulation, the state of the system is described by a distribution over the set of all subsets of the possible labels. That is, each subset A of labels has assigned to it a number representing a probability that the subset of possible labels which are still possible based on the evidence is precisely A (see, e.g., [28]). Implicit in this model is the notion that an incremental piece of evidence carries a certain amount of weight or confidence, and distinguishes a subset of possibilities. Evidence may point to a single inference among the set of labels, or may point to a subset of the alternatives. Further, the Dempster/Shافر theory insists that no mass is placed on the empty set, reflecting the assumption that the label set is exhaustive, so that at least one label must be correct.

As evidence is gained, masses are updated according to a combination formula. The effect of an incremental bit of information pointing to a particular subset A is to transfer partial mass from sets to subsets defined by intersection with A . However, mass moved from subsets that are disjoint from A to the empty set is redistributed evenly among all other subsets. Thus, new evidence typically concentrates mass in low-order subsets, moving mass into subsets, except that mass directed to the empty set is recirculated to all nonempty subsets. The combination formula is commutative and associative, so a succession of incremental changes can be combined into a single state that can be regarded as a non-primitive updating element.

Shafer defines the *belief* on a subset of possibilities A to be the sum of the masses which are applied to subsets of A . This quantity represents a belief in the statement "The truth lies in A ", and corresponds to the "lower probability" in Dempster's formulation. The highest degree of support that the evidence provides for a subset A is the amount of mass that can move to a subset of A , and thus is a sum of masses on subsets that meet A . These values, called "plausibilities", are the "upper probabilities" defined by Dempster. Finally, Shafer defines the "commonality numbers" of a subset A as the sum of masses on subsets which contain A .

Commonality numbers represent the total amount of mass that is available to move to the entire subset A . Shafer shows that the mass function, belief function, plausibility numbers, and commonality numbers are all equivalent formulations of a body of evidence, and that each can be derived from any other [11].

The Dempster/Shافر theory of evidence differs from Bayesian analysis in several important respects. First, beliefs are applied not only to the singleton labels, but also to sets of labels by a non-additive set function. This increase in the dimensionality of the space of states permits distinctions in the different types of evidence that can be represented. For example, lack of evidence can be represented in the theory of evidence by withholding mass to the entire set of possibilities, whereas conflicting evidence is denoted by placing large mass on disjoint subsets. Second, incremental evidence can take a fairly complex form, indicating a subset of possibilities without expressing preferences within the subset. Finally, the combination operation is considerably more complex than termwise multiplication of probabilities, unless all mass is concentrated in singleton subsets. In this way, the combination formula seems to extend Bayesian updating.

Recently, Kyburg [3] has shown how to view the Dempster/Shافر theory of evidence in terms of a collection of probabilistic opinions over the label set. His viewpoint is similar to the one here, in that a state is represented by a set of opinions. However, whereas we view the beliefs as statistics, Kyburg interprets these numbers as extrema over the collection of experts. Accordingly, although he contrasts the combination formula to Bayesian updating on the set of probabilistic opinions, the two viewpoints are different.

In general, the viewpoint that masses and other numbers are assigned to subsets of labels obscures the statistical basis on which the upper and lower probability analysis is based. In Dempster's earlier work [4], however, the set-valued functions are defined over measure spaces, which can each be viewed as a probability space yielding subsets of labels for each sample. In this paper, we return to the earlier Dempster model of measures on a measure space, and relate those notions to spaces of "experts" with opinions expressed as subsets of possibilities. This portion of our formulation appears completely and in greater generality in [4], although we hope to make the connection to pragmatic application issues more explicit.

This paper has three main points. First, we show that the combination rule for the Dempster/Shافر theory of evidence may be simplified by omitting the normalization term. We next point out that the individual pairs of experts involved in the combination formula can be regarded as performing Bayesian updating. Finally, we present extensions to the theory, based on allowing experts to express probabilistic opinions and assuming that the logarithms of experts' opinions over the set of labels are multinormally distributed.

2. The Rule of Combination and Normalization

The set of possible outcomes, or labelings, will be denoted in this paper by Λ . This set is the "frame of discernment", and in other works has been denoted, variously, by Ω , Θ , or S . For convenience, we will assume that Λ is a finite set with n elements, although the framework could easily be extended to continuous label sets. More importantly, we will assume that Λ represents a set of states that are mutually exclusive and exhaustive. If Λ is not initially exhaustive, it can easily be made so by

A Statistical Viewpoint on the Theory of Evidence

including an additional label denoting "none of the above." If Λ is not mutually exclusive, it can be made so by replacement with its power set (i.e., the set of all subsets), so that each subset represents the occurrence of exactly that subset of labels, excluding all other labels. Of course, replacing Λ by its power set is perilous, in that it will greatly expand the cardinality of the label set. For practical applications, the implementer is more likely to want to replace Λ by the set of all plausible subsets describing a valid configuration.

An element (or state of belief) in the theory of evidence is represented by a probability distribution over the power set of Λ , $P(\Lambda)$. That is, a state m is

$$m : P(\Lambda) \rightarrow [0, 1],$$

$$\sum_{A \in \Lambda} m(A) = 1.$$

There is an additional proviso that is typically applied, namely that every state m satisfies

$$m(\emptyset) = 0.$$

Section 3.2 introduces a plausible interpretation for the quantities comprising a state.

A state is updated by combination with new evidence, or information, which is presented in the form of another state. Thus given a current state m_1 , and another state m_2 , a combination of the two states is defined to yield a state $m_1 \oplus m_2$ given by

$$(m_1 \oplus m_2)(A) = \frac{\sum_{B \cap C = A} m_1(B)m_2(C)}{1 - \sum_{B \cap C = \emptyset} m_1(B)m_2(C)} \text{ if } A \neq \emptyset, \quad (1a)$$

and

$$(m_1 \oplus m_2)(\emptyset) = 0.$$

This is the so called "Dempster Rule of Combination." Note that the resulting function m is a probability mass due to the normalization factor, and that $(m_1 \oplus m_2)(\emptyset) = 0$ by definition.

The problem with this definition is that the denominator in (1a) might be zero, so that $(m_1 \oplus m_2)(A)$ is undefined. That is, there exist pairs m_1 and m_2 such that the combination of m_1 and m_2 is not defined. This, of course, is not a very satisfactory situation for a binary operation on a space. The solution which is frequently taken is to avoid combining such elements. An alternative is to add an additional element m_0 to the space:

$$m_0(A) = 0 \text{ for } A \neq \emptyset,$$

$$m_0(\emptyset) = 1.$$

Note that this additional element does not satisfy the condition $m(\emptyset) = 0$. Then define, as a special case,

$$m_1 \oplus m_2 = m_0 \text{ if } \sum_{B \cap C = \emptyset} m_1(B)m_2(C) = 1. \quad (1b)$$

The binary operation is then defined for all pairs m_1, m_2 . The special element m_0 is an absorbent state, in the sense that $m_0 \oplus m = m \oplus m_0 = m_0$ for all states m .

This space has an identity element. The identity state, m_I , represents complete ignorance, in that combination with it yields no change, (i.e., $m_I \oplus m = m \oplus m_I = m$, for all states m). This state places full mass on the subset which is all of Λ ,

$$m_I(\Lambda) = 1$$

$$m_I(A) = 0 \text{ for } A \neq \Lambda.$$

Definition 1: We define (\mathcal{M}, \oplus) , the *space of belief states*, by

$$\mathcal{M} = \{m : \mathcal{P}(\Lambda) \rightarrow \mathbb{R}^+ \cup \{0\} \mid \sum_{A \subseteq \Lambda} m(A) = 1, m(\emptyset) = 0\} \cup \{m_0\},$$

and define \oplus by (1a) when the denominator in (1a) is nonzero, and by (1b) otherwise. ■

The set \mathcal{M} , together with the combination operation \oplus , constitutes a *monoid*, since the binary operation is closed and associative, and there is an identity element.¹ In fact, the binary operation is commutative, so we can say that the space is an abelian monoid.

Still, because of the normalization and the special case in the definition of \oplus , the monoid \mathcal{M} is both ugly and cumbersome. It makes better sense to dispense with the normalization. We have

Definition 2: We define (\mathcal{M}', \oplus') , the *space of unnormalized belief states*, by

$$\mathcal{M}' = \{m : \mathcal{P}(\Lambda) \rightarrow \mathbb{R}^+ \cup \{0\} \mid \sum_{A \subseteq \Lambda} m(A) = 1\}$$

without the additional proviso, and set

$$(m_1 \oplus' m_2)(A) = \sum_{B \cap C = A} m_1(B) \cdot m_2(C) \quad \forall A \subseteq \Lambda \quad (2)$$

for all pairs $m_1, m_2 \in \mathcal{M}'$. ■

One can verify that $m_1 \oplus' m_2 \in \mathcal{M}'$, and that \oplus' is associative and commutative. Further, the same element m_I defined above is also in \mathcal{M}' , and is an identity. Thus \mathcal{M}' is also an abelian monoid. Clearly, \mathcal{M}' is a more attractive monoid than \mathcal{M} .

We define a transformation V mapping \mathcal{M}' to \mathcal{M} by the formulas

$$(Vm)(A) = \frac{m(A)}{1 - m(\emptyset)}, \quad (3)$$

$$(Vm)(\emptyset) = 0$$

if $m(\emptyset) \neq 1$, and

$$Vm = m_0$$

¹A structure with a closed associative binary operation is sometimes call a *semigroup*, so that the space in question is an abelian semigroup with an identity.

otherwise.

A computation shows that V preserves the binary operation; i.e.,

$$V(m_1 \oplus' m_2) = V(m_1) \oplus V(m_2).$$

Thus V is a *homomorphism*.² Further, V is *onto*, since for $m \in \mathcal{M}$, the same m is in \mathcal{M}' , and $Vm = m$. The algebraic terminology is that V is an *epimorphism* of monoids, a fact which we record in

Lemma 1: V maps homomorphically from (\mathcal{M}', \oplus') onto (\mathcal{M}, \oplus) . ■

A "representation" is a term that refers to a map that is an epimorphism of structures. Intuitively, such a map is important because it allows us to consider combination in the space formed by the range of the map as combinations of preimage elements. Lemma 1 will eventually form a small part of a representation to be defined in the next section. In the case in point, however, if it is required to combine elements in \mathcal{M} , one can perform the combinations in \mathcal{M}' , and project to \mathcal{M} by V after all of the combinations are completed. Since combinations in \mathcal{M}' are much cleaner, this is a potentially useful observation. In terms of the Dempster/Shافر theory of evidence, this result says that the normalization in the combination formula is essentially irrelevant, and that combining can be handled by Equation (2). Specifically, given a sequence of states in \mathcal{M} to be combined, say m_1, m_2, \dots, m_k , we can regard these states as elements in \mathcal{M}' . Since each m_i satisfies $m_i(\emptyset) = 0$, they each satisfy $Vm_i = m_i$. Thus

$$V(m_1 \oplus' m_2 \oplus' \dots \oplus' m_k) = Vm_1 \oplus \dots \oplus Vm_k = m_1 \oplus \dots \oplus m_k,$$

which says that it suffices to compute the combinations using \oplus' (Equation (2)), and then project by V (Equation (3)). Of course, the final projection is necessary only if we absolutely insist on a result in \mathcal{M} . If any more combining is to be done, or if we are reasonably broad-minded, intermediate results can be interpreted directly as elements in \mathcal{M}' .

3. Spaces of Opinions of Experts

In this section, we introduce two new spaces, based on the opinions of sample spaces of experts, and discuss the evaluation of statistics of experts opinions. Finally, we interpret the combination rules in these spaces as being a form of Bayesian updating. In the following section we will show that these spaces also map homomorphically onto the space of belief states.

3.1. Opinions of Experts

We consider a set \mathcal{E} of "experts", together with a map μ giving a weight or strength for each expert. It is convenient to think of \mathcal{E} as a large but finite set, although the essential restriction is that \mathcal{E} should be a measure space. Each expert $\omega \in \mathcal{E}$ maintains a list of possible labels: Dempster uses the notation $\Gamma(\omega)$ for this

²Strictly speaking, this merely shows that V is a homomorphism of semigroups; it is not hard to show that V maps the identity to the identity, which it must since it is onto, and thus it is also a homomorphism of monoids.

subset; i.e., $\Gamma(\omega) \subseteq \Lambda$. Here we will assume that each expert ω has more than just a subset of possibilities $\Gamma(\omega)$, but also a *probabilistic opinion* p_ω defined on Λ satisfying

$$p_\omega(\lambda) \geq 0, \forall \lambda \in \Lambda$$

$$p_\omega(\lambda) > 0 \text{ iff } \lambda \in \Gamma(\omega),$$

and

$$\left(\sum_{\lambda \in \Lambda} p_\omega(\lambda) = 1 \text{ or } p_\omega(\lambda) = 0 \forall \lambda \right), \forall \omega \in \mathcal{E}.$$

As suggested by the notation, $p_\omega(\lambda)$ represents expert ω 's assessment of the probability of occurrence of the label λ . If an expert ω believes that a label λ is possible, i.e., $\lambda \in \Gamma(\omega)$, then the associated probability estimate $p_\omega(\lambda)$ will be nonzero. Conversely, if ω thinks that λ is impossible ($\lambda \notin \Gamma(\omega)$), then $p_\omega(\lambda) = 0$. We also include the possibility that expert ω has no opinion which is indicated by the special element $p_\omega = 0$. This state is included in order to ensure that the binary operation, to be defined later, is closed. We denote the collection of maps $\{p_\omega \mid \omega \in \mathcal{E}\}$ by P .

It will turn out that the central point in the theory of evidence is that the $p_\omega(\lambda)$ data is used only in terms of test for zero. Specifically, we set

$$x_\omega(\lambda) = \begin{cases} 1 & \text{if } p_\omega(\lambda) > 0 \\ 0 & \text{if } p_\omega(\lambda) = 0. \end{cases} \quad (4)$$

Note that x_ω is the characteristic function of the set $\Gamma(\omega)$ over Λ , i.e., $x_\omega(\lambda) = 1$ iff $\lambda \in \Gamma(\omega)$. The collection of all x_ω 's will be denoted by X , and will be called the *boolean opinions* of the experts \mathcal{E} .

If we regard the space of experts \mathcal{E} as a sample space, then each $x_\omega(\lambda)$ can be regarded as a sample of a random (boolean) variable $x(\lambda)$. In a similar way, the $p_\omega(\lambda)$'s are also samples of random variables $p(\lambda)$. The state of the system will be defined by statistics on the set of random variables $\{x(\lambda)\}_{\lambda \in \Lambda}$. These statistics are measured over the space of experts. If all experts have the same opinion, then the state should describe that set of possibilities, and the fact that there is a unanimity of opinion. If there is a divergence of opinions, the state should record the fact.

To compute statistics, we will simply sum the weights of experts in subsets of \mathcal{E} . If the experts have equal weights, this is equivalent to counting the number of experts. In general, we will sum the weights of experts in a subset \mathcal{F} , and denote the result by $\mu(\mathcal{F})$. Thus μ is in fact a measure on \mathcal{E} , although it is completely determined by the weights of the individual experts $\mu(\{\omega\})$ for $\omega \in \mathcal{E}$. (We are assuming that \mathcal{E} is finite.) That is,

$$\mu(\mathcal{F}) = \sum_{\omega \in \mathcal{F}} \mu(\{\omega\}).$$

It is important to observe that these measures are evaluated on subsets of experts, and not on the subsets of Λ . The $m(A)$ values which show up in Shafer's work are applied to subsets of the frame of discernment Λ , but are related to the measures μ defined on subsets of experts, as we will presently show. The measures μ show up in Dempster's original work on upper and lower probabilities, however, and are the basis for our presentation that follows. In fact, Dempster treats a more

general case where μ can be a measure defined on a Borel class of subsets of an infinite space of experts \mathcal{E} . For our purposes, it suffices to consider measures on finite sets of experts. It is nearly sufficient to consider nothing more than counting measures on finite sets of experts (experts equally weighted), although we defer an explanation of this point until the end of the Section 4.

We are now ready to introduce the spaces which we will term "opinions of experts." The central point is that the set of labels Λ is fixed, but that the set of experts \mathcal{E} can be different for distinct elements in these spaces. For the first space, we also require a fixed set of positive constants κ_λ , one for each label.

Definition 3: Let $K = \{\kappa_\lambda\}$ be a set of positive constants indexed over the label set Λ . The space of probabilistic opinions of experts $(\mathcal{N}, K, \otimes)$, is defined by

$$\mathcal{N} = \{(\mathcal{E}, \mu, P) \mid \#\mathcal{E} < \infty, \mu \text{ is a measure on } \mathcal{E}, P = \{p_\omega\}_{\omega \in \mathcal{E}},$$

$$p_\omega: \Lambda \rightarrow [0, 1] \forall \omega, \text{ and } \forall \omega, \sum_{\lambda \in \Lambda} p_\omega(\lambda) = 1 \text{ or } p_\omega = 0 \}.$$

As noted earlier, the requirement that $\#\mathcal{E} < \infty$ is for clarity of presentation; Dempster defines the space \mathcal{N} in a more general setting.

We define a binary operations on \mathcal{N} as follows. Given $(\mathcal{E}_1, \mu_1, P_1)$ and $(\mathcal{E}_2, \mu_2, P_2)$ elements in \mathcal{N} , define

$$(\mathcal{E}, \mu, P) = (\mathcal{E}_1, \mu_1, P_1) \otimes (\mathcal{E}_2, \mu_2, P_2)$$

by

$$\mathcal{E} = \mathcal{E}_1 \times \mathcal{E}_2 = \{(\omega_1, \omega_2) \mid \omega_1 \in \mathcal{E}_1, \omega_2 \in \mathcal{E}_2\},$$

$$\mu(\{(\omega_1, \omega_2)\}) = \mu_1(\{\omega_1\}) \cdot \mu_2(\{\omega_2\}),$$

and

$$P = \{p_{(\omega_1, \omega_2)}\}_{(\omega_1, \omega_2) \in \mathcal{E}},$$

$$p_{(\omega_1, \omega_2)}(\lambda) = \frac{p_{\omega_1}^{(1)}(\lambda) p_{\omega_2}^{(2)}(\lambda) [\kappa_\lambda]^{-1}}{\sum_{\lambda'} p_{\omega_1}^{(1)}(\lambda') p_{\omega_2}^{(2)}(\lambda') [\kappa_{\lambda'}]^{-1}},$$

providing the denominator is nonzero, and

$$p_{(\omega_1, \omega_2)} = 0$$

otherwise. Here, $P_i = \{p_{\omega_i}^{(i)}\}_{\omega_i \in \mathcal{E}_i}$ for $i=1, 2$, and the κ_λ 's are a fixed set of positive constants defined for $\lambda \in \Lambda$. ■

To interpret this combining operation, consider two sets of experts \mathcal{E}_1 and \mathcal{E}_2 , with each set of experts expressing opinions in the form of P_1 and P_2 . We form a new set of experts, which is simply the set of all committees of two, consisting of one expert from \mathcal{E}_1 , and another from \mathcal{E}_2 . In each of the committees, the members confer to determine a consensus opinion. In Section 3.3, we will see how to interpret the formulas as Bayesian combination (where κ_λ is the prior probability on λ). And in the following section we will show that this space maps homomorphically

onto the belief spaces. Finally, if as in Dempster [4], we only regard the opinions of these experts in terms of a test for zero (i.e. disregarding the strength of nonzero opinions), we arrive at yet another space. A depiction of the combination of two Boolean opinions is shown in Figure 1.

Definition 4: The *space of boolean opinions of experts*, (\mathcal{N}', \cdot) , is defined similarly:

$$\mathcal{N}' = \{(\mathcal{E}, \mu, X) \mid \# \mathcal{E} < \infty, \mu \text{ is a measure on } \mathcal{E},$$

$$X = \{x_\omega\}_{\omega \in \mathcal{E}}, \quad x_\omega : \Lambda \rightarrow \{0, 1\} \forall \omega\}.$$

If $(\mathcal{E}_1, \mu_1, X_1)$ and $(\mathcal{E}_2, \mu_2, X_2)$ are elements in \mathcal{N}' , define their product

$$(\mathcal{E}, \mu, X) = (\mathcal{E}_1, \mu_1, X_1) \odot (\mathcal{E}_2, \mu_2, X_2)$$

by

$$\mathcal{E} = \mathcal{E}_1 \times \mathcal{E}_2 = \{(\omega_1, \omega_2) \mid \omega_1 \in \mathcal{E}_1, \omega_2 \in \mathcal{E}_2\}$$

$$\mu(\{(\omega_1, \omega_2)\}) = \mu_1(\{\omega_1\}) \cdot \mu_2(\{\omega_2\}),$$

and

$$X = \{x_{(\omega_1, \omega_2)}\}_{(\omega_1, \omega_2) \in \mathcal{E}},$$

$$x_{(\omega_1, \omega_2)}(\lambda) = x_{\omega_1}^{(1)}(\lambda) \cdot x_{\omega_2}^{(2)}(\lambda),$$

where $X_i = \{x_{\omega_i}^{(i)} \mid \omega_i \in \mathcal{E}_i\}$, for $i = 1, 2$. ■

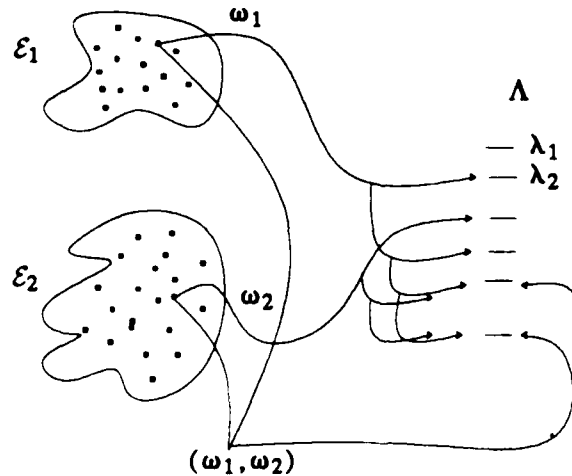


Figure 1. A depiction of the combination of two boolean opinions of two experts, as is present in combinations in \mathcal{N}' , yielding a consensus opinion by the element in the product set of experts formed by the committee of two.

3.2. Statistics of Experts

For a given subset $A \subseteq \Lambda$, the characteristic function χ_A is defined by

$$\chi_A(\lambda) = \begin{cases} 0 & \text{if } \lambda \notin A \\ 1 & \text{if } \lambda \in A. \end{cases}$$

Equality of two functions defined on Λ means, of course, that the two functions agree for all $\lambda \in \Lambda$. That is, $x_\omega = \chi_A$ means

$$x_\omega(\lambda) = \chi_A(\lambda) \quad \forall \lambda \in \Lambda,$$

which is the same thing as saying $\Gamma(\omega) = A$.

Given a space of experts \mathcal{E} and the boolean opinions X , we define

$$\bar{m}(A) = \frac{\mu\{\omega \in \mathcal{E} \mid x_\omega = \chi_A\}}{\mu\{\mathcal{E}\}} \quad (5)$$

for every subset $A \subseteq \Lambda$. It is possible to view the values as probabilities on the random variables $\{x(\lambda)\}$. We endow the elements of \mathcal{E} with the prior probabilities $\mu(\{\omega\})/\mu(\mathcal{E})$, and say that the probability of an event involving a combination of the random variables $x(\lambda)$'s over the sample space \mathcal{E} is the probability that the event is true for a particular sample, where the sample is chosen at random from \mathcal{E} with the sampling distribution given by the prior probabilities. This is equivalent to saying

$$\text{Prob}_{\mathcal{E}}(\text{Event}) = \frac{\mu(\{\omega \in \mathcal{E} \mid \text{Event is true for } \omega\})}{\mu\{\mathcal{E}\}}.$$

With this convention, we see that

$$\bar{m}(A) = \text{Prob}_{\mathcal{E}}(x(\lambda) = \chi_A(\lambda) \text{ for all } \lambda).$$

In fact, all of the priors and joint statistics of the $x(\lambda)$'s are determined by the full collection of $\bar{m}(A)$ values. For example,

$$\text{Prob}(x(\lambda_0) = 1) = \sum_{\{A \mid \lambda_0 \in A\}} \bar{m}(A)$$

and

$$\text{Prob}(x(\lambda_0) = 1 \text{ and } x(\lambda_1) = 1) = \sum_{\{A \mid \lambda_0, \lambda_1 \in A\}} \bar{m}(A).$$

Further, the full set of values $\bar{m}(A)$ for $A \subseteq \Lambda$ defines an element $\bar{m} \in \mathcal{M}'$. To see this, it suffices to check that $\sum \bar{m}(A) = 1$, which amounts to observing that for every ω , $x_\omega = \chi_A$ for some $A \subseteq \Lambda$.

Recalling the definition of V (Equation (3)), we may also consider the numbers $(V\bar{m})(A)$. These values can also be interpreted as probabilities, providing we define probability in a way which ignores experts who give no possibilities, and providing there are some experts who give some possibilities, (i.e., $m(\emptyset) \neq 1$). Then for $A \neq \emptyset$,

$$m(A) = (V\bar{m})(A) = \frac{\bar{m}(A)}{1 - \bar{m}(\emptyset)}$$

is the probability that a randomly chosen expert ω will state that the subset of

possibilities is precisely A conditioned on the requirement that the expert gives at least one possibility.

Under the assumptions that $A \neq \emptyset$, $\bar{m}(\emptyset) \neq 1$, and that probability is measured over the set of experts expressing an opinion $\mathcal{E}' = \{\omega | x_\omega \neq 0\}$, many of the quantities in the theory of evidence can be interpreted in terms of familiar statistics on the $x(\lambda)$'s. For example, the belief on a set A ,

$$\text{Bel}(A) = \sum_{B \subseteq A} m(B)$$

is simply the joint probability

$$\text{Bel}(A) = \text{Prob}(x(\lambda) = 0 \text{ for } \lambda \notin A).$$

Note that the prior probabilities on the experts in \mathcal{E}' are given by $\mu(\{\omega\})/\mu(\mathcal{E}')$. The denominator in these priors is nonzero due to the assumption that $\bar{m}(\emptyset) \neq 1$.

In a similar way, plausibility values

$$\text{Pl}(A) = \sum_{B \cap A \neq \emptyset} m(B) = 1 - \text{Bel}(\bar{A})$$

can be interpreted as disjunctive probabilities

$$\text{Pl}(A) = \text{Prob}(x(\lambda) = 1 \text{ for some } \lambda \in A).$$

The beliefs and plausibilities are the lower and upper probabilities as defined by Dempster. The commonality values

$$Q(A) = \sum_{A \subseteq B} m(B)$$

are joint probabilities:

$$Q(A) = \text{Prob}(x(\lambda) = 1 \text{ for } \lambda \in A).$$

To recapitulate, we have defined a mapping from P values to X values, and then transformations from X to \bar{m} and m values. The resulting element m , which contains statistics on the X variables, is an element in the space of belief states \mathcal{M} of the of the Dempster/Shafar theory of evidence (Section 2).

3.3. Bayesian Interpretation

We now interpret the manner in which pairs of experts achieve a consensus opinion. We will show that the combination formulas given for \mathcal{N} and \mathcal{N}' are consistent with a Bayesian interpretation. Our treatment is standard.

We first consider the combination of $(\mathcal{E}_1, \mu_1, P_1)$ and $(\mathcal{E}_2, \mu_2, P_2)$ in \mathcal{N} . We assume that the experts in \mathcal{E}_j have available to them information s_j . Note that all experts in a given set of experts share the same information. The information s_j consists of boolean predicates constituting evidence about the labeling situation. For example, in a medical diagnosis application, s_j might consist of a statements about the presence or absence of a set of symptoms. Each set of experts \mathcal{E}_j deals with a different set of symptoms.

A Statistical Viewpoint on the Theory of Evidence

In general, the information s_j is the result of a set of tests having boolean outcomes. We could write $s_j = f_j(\sigma)$, where f_j represents the tests, and σ is the current situation which is an element in some sample space of labeling problems $\sigma \in \Sigma$. Assuming Σ is also a measure space, there are prior probabilities on the information coefficients:

$$\text{Prob}(s_j) = \text{Prob}(f_j(\sigma) = s_j).$$

There are also prior probabilities on the true label $\lambda(\sigma)$ for labeling situation σ , given by

$$\text{Prob}(\lambda) = \text{Prob}(\lambda(\sigma) = \lambda).$$

Note that these probabilities are not measured over the space of experts \mathcal{E} , but instead are measured over the collection of instances Σ of the labeling problem. For example, in a medical diagnosis domain, Σ might represent the set of all patients.

For $j=1,2$, we will suppose that $p_{\omega_j}^{(j)}(\lambda)$ represents expert ω_j 's estimate of

$$\text{Prob}(\lambda | s_j),$$

the probability (over Σ) that $\lambda(\sigma) = \lambda$ conditioned on $f_j(\sigma) = s_j$. The "expert" (ω_1, ω_2) should then estimate $\text{Prob}(\lambda | s_1, s_2)$, which is the probability that $\lambda(\sigma) = \lambda$ given that $f_1(\sigma) = s_1$ and $f_2(\sigma) = s_2$, thus combining the two bodies of evidence seen by the two experts in that committee. This committee proceeds as follows:

Bayes' formula implies that

$$\text{Prob}(\lambda | s_1, s_2) = \frac{\text{Prob}(\lambda) \cdot \text{Prob}(s_1, s_2 | \lambda)}{\text{Prob}(s_1, s_2)} = \frac{\text{Prob}(\lambda) \cdot \text{Prob}(s_1 | \lambda) \cdot \text{Prob}(s_2 | s_1, \lambda)}{\text{Prob}(s_1, s_2)}.$$

Applying Bayes' formula to $\text{Prob}(s_1 | \lambda)$, this becomes

$$\frac{\text{Prob}(s_1)}{\text{Prob}(s_1, s_2)} \cdot \text{Prob}(\lambda | s_1) \cdot \text{Prob}(s_2 | s_1, \lambda) \quad (6)$$

At this point that we assume that

$$\text{Prob}(s_2 | s_1, \lambda) = \text{Prob}(s_2 | \lambda). \quad (7)$$

Using this assumption, we obtain by combining (6) and (7), and applying Bayes' formula to $\text{Prob}(s_2 | \lambda)$,

$$\text{Prob}(\lambda | s_1, s_2) = c(s_1, s_2) \cdot \frac{\text{Prob}(\lambda | s_1) \cdot \text{Prob}(\lambda | s_2)}{\text{Prob}(\lambda)}, \quad (8)$$

where $c(s_1, s_2)$ is a constant independent of λ . Using Equation (8), expert (ω_1, ω_2) estimates that

$$p_{(\omega_1, \omega_2)}(\lambda) = c(s_1, s_2) \cdot \frac{p_{\omega_1}^{(1)}(\lambda) p_{\omega_2}^{(2)}(\lambda)}{\kappa_\lambda}, \quad (9)$$

based on the independence assumption (7), where $\kappa_\lambda = \text{Prob}(\lambda)$. Since the left hand side of this equation should sum to 1 over λ , we have that

$$c(s_1, s_2) = \frac{1}{\sum_{\lambda'} p_{\omega_1}^{(1)}(\lambda') p_{\omega_2}^{(2)}(\lambda') [\kappa_{\lambda'}]^{-1}}, \quad (10)$$

unless, of course, this denominator is zero, in which case we resort to setting $p(\omega_1, \omega_2) = 0$. Combining (9) and (10) gives the combination formula given in Definition 3. Thus, we have shown that combination in \mathcal{N} is a form of Bayesian updating of pairs of experts, based on an independence assumption.

To interpret the combination formula of \mathcal{N}' in a Bayesian fashion, a weaker independence assumption suffices. The combination formula can be restated as:

$$x_{(\omega_1, \omega_2)}(\lambda) = 0 \text{ iff } x_{\omega_1}^{(1)}(\lambda) = 0 \text{ or } x_{\omega_2}^{(2)}(\lambda) = 0.$$

Using Bayes' formula, and assuming that all prior probabilities are nonzero, it suffices to show that

$$\text{Prob}(s_1, s_2 | \lambda) = 0 \text{ iff } \text{Prob}(s_1 | \lambda) = 0 \text{ or } \text{Prob}(s_2 | \lambda) = 0.$$

The "if" part follows since

$$\begin{aligned} \text{Prob}(s_1, s_2 | \lambda) &= \text{Prob}(s_1 | \lambda) \cdot \text{Prob}(s_2 | s_1, \lambda) \\ &= \text{Prob}(s_2 | \lambda) \cdot \text{Prob}(s_1 | s_2, \lambda). \end{aligned}$$

The "only if" part becomes our independence assumption, and is equivalent to

$$\text{Prob}(s_1 | \lambda) > 0 \text{ and } \text{Prob}(s_2 | \lambda) > 0 \Rightarrow \text{Prob}(s_1, s_2 | \lambda) > 0. \quad (11)$$

This assumption is implied by our earlier hypothesis (7). However, assumption (11) is more defensible, and is actually all that is needed to regard updating in the space of "boolean opinions of experts," \mathcal{N}' , as Bayesian. Since the Dempster/Shافر theory deals only with the boolean opinions, Equation (11) is the required independence assumption.

4. Equivalence with the Dempster/Shافر Rule of Combination

At this point, we have four spaces with binary operations, namely (\mathcal{N}, \otimes) , (\mathcal{N}', \odot) , (\mathcal{M}', \oplus') , and (\mathcal{M}, \oplus) . We will now show that these four spaces are closely related. It is not hard to show that the binary operation is, in all four cases, commutative and associative, and that each space has an identity element, so that these spaces are abelian monoids. We also have

Definition 5: The map T

$$T: \mathcal{N} \rightarrow \mathcal{N}',$$

with $(\mathcal{E}, \mu, X) = T(\mathcal{E}, \mu, P)$, is given by equation (4), i.e., $x_{\omega}(\lambda) = 1$ iff $p_{\omega}(\lambda) > 0$, and $x_{\omega}(\lambda) = 0$ otherwise. ■

There is another mapping U , given by

Definition 6:

$$U: \mathcal{N}' \rightarrow \mathcal{M}'$$

with $\tilde{m} = U(\mathcal{E}, \mu, X)$ given by equation (5), i.e.,

$$\bar{m}(A) = \mu(\{\omega \in \mathcal{E} | x_\omega = \chi_A\}) / \mu(\{\mathcal{E}\}). \quad \blacksquare$$

We will show that T and U preserve the binary operations. More formally, we show that T and U are homomorphisms of monoids.

Lemma 2: T is a homomorphism from \mathcal{N} onto \mathcal{N}' .

Proof: It is a simple matter to verify that

$$T(\mathcal{E}_1, P_1) \odot T(\mathcal{E}_2, P_2) = T((\mathcal{E}_1, P_1) \otimes (\mathcal{E}_2, P_2)).$$

The essential point, it turns out, is that since the probabilistic opinions are all non-negative,

$$p_{\omega_1}^{(1)}(\lambda) \cdot p_{\omega_2}^{(2)}(\lambda) > 0 \text{ iff } p_{\omega_1}^{(1)}(\lambda) > 0 \text{ and } p_{\omega_2}^{(2)}(\lambda) > 0.$$

T is easily seen to be onto. \blacksquare

Lemma 3: U is a homomorphism of \mathcal{N}' onto \mathcal{M}' .

Proof: Consider $(\mathcal{E}, \mu, X) = (\mathcal{E}_1, \mu_1, X_1) \odot (\mathcal{E}_2, \mu_2, X_2)$. For each $\omega_1 \in \mathcal{E}_1$ and $\omega_2 \in \mathcal{E}_2$, the corresponding $x_{\omega_1}^{(1)}$ and $x_{\omega_2}^{(2)}$ are characteristic functions of subsets of Λ , say χ_B and χ_C respectively. It is clear that

$$x_{\omega_1}^{(1)} \cdot x_{\omega_2}^{(2)} = \chi_A \text{ iff } B \cap C = A.$$

Thus

$$x_{(\omega_1, \omega_2)} = \chi_A \text{ iff } x_{\omega_1}^{(1)} = \chi_B \text{ and } x_{\omega_2}^{(2)} = \chi_C \text{ where } B \cap C = A.$$

So

$$\{(\omega_1, \omega_2) \in \mathcal{E} \mid x_{(\omega_1, \omega_2)} = \chi_A\} = \bigcup_{B \cap C = A} \{\omega_1 \mid x_{\omega_1}^{(1)} = \chi_B\} \times \{\omega_2 \mid x_{\omega_2}^{(2)} = \chi_C\}$$

Since this is a disjoint union, using properties of measures, this gives

$$\mu\{(\omega_1, \omega_2) \in \mathcal{E} \mid x_{(\omega_1, \omega_2)} = \chi_A\} = \sum_{B \cap C = A} \mu_1\{\omega_1 \in \mathcal{E}_1 \mid x_{\omega_1}^{(1)} = \chi_B\} \cdot \mu_2\{\omega_2 \in \mathcal{E}_2 \mid x_{\omega_2}^{(2)} = \chi_C\}.$$

We can divide both sides of this equation by $\mu\{\mathcal{E}\} = \mu_1\{\mathcal{E}_1\} \cdot \mu_2\{\mathcal{E}_2\}$ to obtain

$$\bar{m}(A) = \sum_{B \cap C = A} \bar{m}_1(B) \bar{m}_2(C),$$

where $\bar{m} = U(\mathcal{E}, \mu, X)$, and $\bar{m}_i = U(\mathcal{E}_i, \mu_i, X_i)$, $i=1,2$. Thus

$$U((\mathcal{E}_1, \mu_1, X_1) \odot (\mathcal{E}_2, \mu_2, X_2)) = U(\mathcal{E}_1, \mu_1, X_1) \otimes' U(\mathcal{E}_2, \mu_2, X_2),$$

which is to say that U is a homomorphism.

Finally, we show that U is onto. Recall that there are n elements in Λ , and so there are 2^n different subsets of Λ . For a given mass distribution $\bar{m} \in \mathcal{M}'$, consider a set of 2^n experts \mathcal{E} , with each expert $\omega \in \mathcal{E}$ giving a distinct subset $\Gamma(\omega) \subseteq \Lambda$ as the set of possibilities. If we give expert ω the weight $\mu\{\omega\} = \bar{m}(\Gamma(\omega))$, and set $x_\omega = \chi_{\Gamma(\omega)}$, then it is easy to see that $\bar{m} = U(\mathcal{E}, \mu, X)$. \blacksquare

In the immediately preceding proof that U is onto, we assigned weights to experts. This is the only place where we absolutely require the existence of differential weights on experts. However, if we content ourselves to spaces \mathcal{M}' and \mathcal{M}

containing only rational values for the mass distribution functions (as, for example, is the case in any computer implementation), then the weights can be eliminated, and replaced by counting measure. For in this case, given a rational mass distribution \tilde{m} , we multiply by a common multiple of the denominators of the fractions appearing in the 2^n values of the mass distribution to obtain 2^n integer values in proportion to the given $\tilde{m}(A)$ values. We then construct an element in \mathcal{N}' by replicating each expert the appropriate number of times, given by the integer value corresponding to the subset $A \subseteq \Lambda$ designated by the expert as the subset of possibilities.

Recall from Section 2 that the map $V: \mathcal{M}' \rightarrow \mathcal{M}$ is also a homomorphism. So we can compose the homomorphisms $T: \mathcal{N} \rightarrow \mathcal{N}'$ with $U: \mathcal{N}' \rightarrow \mathcal{M}'$ with $V: \mathcal{M}' \rightarrow \mathcal{M}$ to obtain the following obvious theorem.

Theorem: The map $V \circ U \circ T: \mathcal{N} \rightarrow \mathcal{M}$ is a homomorphism of monoids mapping onto the space of belief states (\mathcal{M}, \oplus) . ■

This theorem provides the justification for the viewpoint that the theory of evidence space \mathcal{M} represents the space \mathcal{N} via the representation $V \circ U \circ T$. The proof follows from the lemmas; since each of the component maps in this representation is an onto homomorphism, the composition also maps homomorphically onto the entire theory of evidence space.

The significance of this result is that we can regard combinations of elements in the theory of evidence as combinations of elements in the space of opinions of experts. For if m_1, \dots, m_k are elements in \mathcal{M} which are to be combined under \oplus , we can find respective preimages in \mathcal{N} under the map $V \circ U \circ T$, and then combine those elements using the operation \otimes in the space of opinions of experts \mathcal{N} . After all combinations in \mathcal{N} are completed, we project back to \mathcal{M} by $V \circ U \circ T$; the result will be the same as if we had combined the elements in \mathcal{M} . The only advantage to this procedure is that combinations in \mathcal{N} are conceptually simpler: there are no funny normalizations, and we can regard the combination as Bayesian updatings on the product space of experts.

5. An Alternative Method for Combining Evidence

With the viewpoint that the theory of evidence is really simply statistics of opinions of experts, we can make certain remarks on the limitations of the theory.

- (1) There is no use of probabilities or degrees of confidence. Although the belief values seem to give weighted results, at the base of the theory experts only say whether a condition is possible or not. In particular, the theory makes no distinction between an expert's opinion that a label is likely or that it is remotely possible.
- (2) Pairs of experts combine opinions in a Bayesian fashion with independence assumptions of the sources of evidence. In particular, dependencies in the sources of information are not taken into account.
- (3) Combinations take place over the product space of experts. It might be more reasonable to have a single set of experts modifying their opinions as new information comes in, instead of forming the set of all committees of mixed pairs.

Both the second and third limitations come about due to the desire to have a combination formula which factors through to the statistics of the experts and is application-independent. The need for the second limitation, the independence assumption on the sources of evidence, is well-known (see, e.g., [14]). Without incorporating much more complicated models of judgements under multiple sources of knowledge, we can hardly expect anything better.

The first objection, however, suggests an alternate formulation which makes use of the probabilistic assessments of the experts. Basically, the idea is to keep track of the density distributions of the opinions in probability space. Of course, complete representation of the distribution would amount to recording the full set of opinions p_ω for all ω . Instead, it is more reasonable to approximate the distribution by some parameterization, and update the distribution parameters by combination formulas.

We present a formulation based on normal distributions of logarithms of updating coefficients. Other formulations are possible. In marked contrast to the Dempster/Shافر formulation, we assume that all opinions of all experts are nonzero for every label. That is, instead of converting opinions into boolean statements by test for zero, we will assume that all the values are nonzero, and model the distribution of their strengths.

A simple rewrite of Equation (8) of Section 3.3 yields

$$\text{Prob}(\lambda | s_1, s_2) = c(s_1, s_2) \cdot \text{Prob}(\lambda) \cdot \frac{\text{Prob}(\lambda | s_1)}{\text{Prob}(\lambda)} \cdot \frac{\text{Prob}(\lambda | s_2)}{\text{Prob}(\lambda)}.$$

This equation depends on an independence assumption, Equation (7). We can iterate this equation to obtain a formula for $\text{Prob}(\lambda | s_1, \dots, s_k)$. In this iteration process, s_1 and s_2 successively take the place of $s_1 \wedge \dots \wedge s_i$ and s_{i+1} respectively, as i increases from 1 to $k-1$. Accordingly, we require a sequence of independence assumptions, which will take the form

$$\text{Prob}(s_{i+1} | s_1 \wedge \dots \wedge s_i, \lambda) = c(s_1, \dots, s_{i+1}) \cdot \text{Prob}(s_i | \lambda)$$

for $i = 1, \dots, k-1$. Under these assumptions, we obtain

$$\text{Prob}(\lambda | s_1, \dots, s_k) = c(s_1, \dots, s_k) \cdot \text{Prob}(\lambda) \cdot \prod_{i=1}^k \frac{\text{Prob}(\lambda | s_i)}{\text{Prob}(\lambda)}.$$

In a manner similar to [29], set

$$L(\lambda | s_i) = \log \left[\frac{\text{Prob}(\lambda | s_i)}{\text{Prob}(\lambda)} \right].$$

(Note, incidentally, that these values are not the so-called "log-likelihood ratios"; in particular, the $L(\lambda | s_i)$'s can be both positive and negative). We then obtain

$$\log[\text{Prob}(\lambda | s_1, \dots, s_k)] = c + \log[\text{Prob}(\lambda)] + \sum_{i=1}^k L(\lambda | s_i),$$

where c is a constant independent of λ (but not of s_1, \dots, s_k).

The consequence of this formula is that if the independence assumptions hold, and if $\text{Prob}(\lambda)$ and $L(\lambda | s_i)$ are known for all λ and i , then the approximate values $\text{Prob}(\lambda | s_1, \dots, s_k)$ can be calculated from

$$\text{Prob}(\lambda | s_1, \dots, s_k) = \frac{\text{Prob}(\lambda) \cdot \exp\left[\sum_{i=1}^k L(\lambda | s_i)\right]}{\sum_{\lambda'} \text{Prob}(\lambda') \exp\left[\sum_{i=1}^k L(\lambda' | s_i)\right]}. \quad (12)$$

Accordingly, we introduce a space which we term "logarithmic opinions of experts." For convenience, we will assume that experts have equal weights. An element in this space will consist of a set of experts \mathcal{E}_i , and a collection of opinions $Y_i = \{y_{\omega}^{(i)}\}_{\omega \in \mathcal{E}_i}$. Each $y_{\omega}^{(i)}$ is a map, and the component $y_{\omega}^{(i)}(\lambda)$ represents expert ω 's estimate of $L(\lambda | s_i)$:

$$y_{\omega}^{(i)} : \Lambda \rightarrow \mathbb{R}, \quad y_{\omega}^{(i)}(\lambda) \approx L(\lambda | s_i).$$

Note that the experts in \mathcal{E}_i all have knowledge of the information s_i , and that the estimated logarithmic coefficients $L(\lambda | s_i)$ can be positive or negative. In fact, since the experts do not necessarily have precise knowledge of the value of $\text{Prob}(\lambda)$, but instead provide estimates of log's of ratios, the estimates can lie in an unbounded range.

In analogy with our map to a statistical space (Section 3.2), we can define a space which might be termed the "parameterized statistics of logarithmic opinions of experts." Elements in this space will consist of pairs (\bar{u}, C) , where \bar{u} is in \mathbb{R}^n and C is a symmetric n by n matrix. We next describe how to project from the space of logarithmic opinions to the space of parameterized statistics.

Let us suppose that for a set of experts \mathcal{E} , and for $\Lambda = \{\lambda_1, \dots, \lambda_n\}$, the n -vectors composed of the logarithmic opinions $\bar{y}_{\omega} \in \mathbb{R}^n$, $\bar{y}_{\omega} = (y_{\omega}(\lambda_1), \dots, y_{\omega}(\lambda_n))$, are approximately (multi-) normally distributed. Thus we model the distribution of the random vector $\bar{y} = (y(\lambda_1), \dots, y(\lambda_n))$ by the density function

$$m(\bar{y}) = \frac{1}{(2\pi)^{n/2} \sqrt{\det C}} \exp((\bar{y} - \bar{u})^T C^{-1} (\bar{y} - \bar{u})), \quad \bar{y} \in \mathbb{R}^n,$$

where $\bar{u} \in \mathbb{R}^n$ is the mean of the distribution, and C is the n by n covariance matrix. That is, in terms of the expectation operator $E\{\cdot\}$ on random variables over the sample space \mathcal{E} ,

$$\bar{u} = (u_1, \dots, u_n),$$

$$u_i = E\{y(\lambda_i)\},$$

and for $C = (c_{ij})$,

$$c_{ij} = E\{(y(\lambda_i) - u_i)(y(\lambda_j) - u_j)\}.$$

These measurements of the statistics of the $y(\lambda)$'s can be made regardless of the true distributions. The accuracy of the model depends on the degree to which the multinormal distribution assumption is valid.

Next we discuss combination formulas in both spaces. Suppose (\mathcal{E}_i, Y_i) , $i = 1, 2$, are two elements in the space of logarithmic opinions, each describing a sample space of experts together with opinions. Since according to Equation (12), the logarithmic opinions add, we define the combination of the two elements by (\mathcal{E}, Y) , where

A Statistical Viewpoint on the Theory of Evidence

$$\mathcal{E} = \mathcal{E}_1 \times \mathcal{E}_2,$$

$$Y = \{y(\omega_1, \omega_2)\}_{(\omega_1, \omega_2) \in \mathcal{E}},$$

$$y(\omega_1, \omega_2)(\lambda) = y_{\omega_1}^{(1)}(\lambda) + y_{\omega_2}^{(2)}(\lambda).$$

To consider combinations in the space of statistics, let $m_i(\bar{y})$ be the density function over \mathbb{R}^n for the random vector $\bar{y}^{(i)}$ over the sample space \mathcal{E}_i , $i = 1, 2$. Assume that each m_i is a multinormal distribution, associated with a mean vector $\bar{u}^{(i)}$ and a covariance $C^{(i)}$. In order that the projection to the space of statistics be a homomorphism, the definition of combination in the space of statistics should respect the true statistics of the combined opinions. The density function $m(\bar{y})$ for the combination $\bar{y}_{(\omega_1, \omega_2)}$, $(\omega_1, \omega_2) \in \mathcal{E}$, is given by

$$m(\bar{y}) = \int_{\mathbb{R}^n} m_1(\bar{y}') m_2(\bar{y} - \bar{y}') d\bar{y}'.$$

Notice that this is the point where we use the fact that the logarithmic opinions add under combination.

Projecting to the space of statistics, we discover the advantage of modeling the distributions by normal functions. Namely, since the convolution of a Gaussian by a Gaussian is once again a Gaussian, we define the combination formula

$$(\bar{u}^{(1)}, C^{(1)}) \oplus (\bar{u}^{(2)}, C^{(2)}) = (\bar{u}^{(1)} + \bar{u}^{(2)}, C^{(1)} + C^{(2)}).$$

That is, since m_1 and m_2 are multinormal distributions, their convolution is also multinormal with mean and covariance which are the sums of the contributing means and covariances. (This result is easily proven using Fourier transforms.) An extension to the case where \mathcal{E}_1 and \mathcal{E}_2 have nonequal total weights is straightforward.

Having defined combination in the space of statistics, one must show that the transformation from the space of opinions to the space of statistics is a homomorphism, even when the logarithmic opinions are not truly normally-distributed. This is easily done, since the means and covariances of the sum of two random vectors are the sums of the means and covariances of the two random vectors.

To interpret a state (\bar{u}, C) in the space of parameterized statistics, we must remember the origin of the logarithmic-opinion values. Specifically, after k updating iterations combining information s_1 through s_k , the updated vector $\bar{y} = (y_1, \dots, y_n) \in \mathbb{R}^n$ is an estimate of the sum of the logarithmic coefficients,

$$y_j \approx \sum_{i=1}^k L(\lambda | s_i).$$

According to Equation (12), the a posteriori probabilities can then be calculated from this estimate (providing the priors $\text{Prob}(\lambda)$'s are known). In particular, the a posteriori probability of a label λ_j is high if the corresponding coefficient $y_j + \log[\text{Prob}(\lambda_j)]$ is large in comparison to the other components $y_j + \log[\text{Prob}(\lambda_j)]$.

Since the state (\bar{u}, C) represents a multinormal distribution in the log-updating space, we can transform this distribution to a density function for a posteriori probabilities. Basically, a label will have a high probability if $u_j + \log[\text{Prob}(\lambda_j)]$ is

relatively large. However, the components of \bar{u} represent the center of the distribution (before bias by the priors). The spread of the distribution is given by the covariance matrix, which can be thought of as defining an ellipsoid in \mathbb{R}^n centered at \bar{u} . The exact equation of the ellipse can be written implicitly as:

$$(\bar{y} - \bar{u})^T C^{-1} (\bar{y} - \bar{u}) = 1.$$

This ellipse describes a "one sigma" variation in the distribution, representing a region of uncertainty of the logarithmic opinions; the distribution to two standard deviations lies in a similar but enlarged ellipse. The eigenvalues of C give the squared lengths of the semi-major axes of the ellipse, and are accordingly proportional to degrees of confidence. The eigenvectors give the directions in which the eigenvalues measure their uncertainty. Bias by the prior probabilities simply adds a fixed vector, with components $\log[\text{Prob}(\lambda_j)]$, to the ellipse, thereby translating the distribution. We seek an axis j such that the components y_j of the vectors y lying in the translated ellipse are relatively much larger than other components of vectors in the ellipse. In this case, the preponderant evidence is for label λ_j .

Clearly, the combination formula is extremely simple. Its greatest advantage over the Dempster/Shافر theory of evidence is that only $O(n^2)$ values are required to describe a state, as opposed to the 2^n values used for a mass distribution in \mathcal{M} . The simplicity and reduction in numbers of parameters has been purchased at the expense of an assumption about the kinds of distributions that can be expected. However, the same assumption allows us to track probabilistic opinions (or actually, the logarithms), instead of converting all opinions into boolean statements about possibilities.

6. Conclusions

We have shown how the theory of evidence may be viewed as a representation of a space of opinions of experts, where opinions are combined in a Bayesian fashion over the product space of experts. (Refer to Figure 2.) By "representation", we mean something very specific — namely, that there is a homomorphism mapping from the space of opinions of experts onto the Dempster/Shافر theory of evidence space. This map fails to be an isomorphism (which would imply equivalence of the spaces) only insofar as it is many-to-one. That is, for each state in the theory of evidence, there is a collection of elements in the space of opinions of experts which all map to the single state. In this way the state in the theory of evidence represents the corresponding collection of elements. In fact, what this collection of elements have in common is that the statistics of the opinions of the experts defined by the element are similar, in terms of the way statistics are measured by the map U .

Furthermore, combination in the space of opinions of experts, as defined in Section 3, leads to combination in the theory of evidence space. This allows us to implement combination in a somewhat simpler manner, since the formulas for combination without the normalization are simpler than the more standard formulas, and also permits us to view combination in the theory of evidence space as the tracking of statistics of opinions of experts as they combine information in a pairwise Bayesian fashion over the product space of experts. Applying a Bayesian interpretation to the updating of the opinions of experts also makes clear the implicit

independence assumptions which must exist in order to combine evidence in the prescribed manner.

From this viewpoint, we can see how the Dempster/Shافر theory of evidence accomplishes its goals. Degrees of support for a proposition, belief, and plausibilities, are all measured in terms of joints and disjunctive probabilities over a set of experts who are naming possible labels given current information. The problem of ambiguous knowledge versus uncertain knowledge, which is frequently described in terms of "withholding belief," can be viewed as two different distributions of opinions. In particular, ambiguous knowledge can be seen as observing high densities of opinions on particular disjoint subsets, whereas uncertain knowledge corresponds to unanimity of opinions, where the agreed upon opinion gives many possibilities. Finally, instead of performing Bayesian updating, a *set* of values are updated in a Bayesian fashion over the product space, which results in non-Bayesian formulas over the space of labels.

In meeting each of these goals, the theory of evidence invokes compromises that we might wish to change. For example, in order to track statistics, it is necessary to model the distribution of opinions. If these opinions are probabilistic assignments over the set of labels, then the distribution function will be too complicated to retain precisely. The Dempster/Shافر theory of evidence solves this problem by simplifying the opinions to boolean decisions, so that each expert's opinion lies in a space having 2^n elements. In this way, the full set of statistics can be specified using 2^n values. We have suggested an alternate method, which retains the probability values in the opinions without converting them into boolean decisions, and requires only $O(n^2)$ values to model the distribution, but fails to retain full information about the distribution. Instead, our method attempts to approximate the distribution of opinions with a Gaussian function.

References

- [1] J. C. Falmagne, "A random utility model for a belief function," *Synthese* 57, pp. 35-48 (1983).
- [2] B.O. Koopman, "The axioms and algebra of intuitive probability," *Ann. Math.* 41, pp. 269-292 (1940). See also "The bases of probability", *Bulletin of the American Math. Society* 46, 1940, p. 763-774.
- [3] Henry E. Kyburg, Jr., "Bayesian and non-bayesian evidential updating," University of Rochester Dept. of Computer Science Tech. Rep. 139 (July, 1985).
- [4] A. P. Dempster, "Upper and lower probabilities induced by a multivalued mapping," *Annals of Mathematical Statistics* 38, pp. 325-339 (1967).
- [5] Peter C. Fishburn, *Decision and Value Theory*, Wiley, New York (1964).
- [6] I. J. Good, "The measure of a non-measurable set," pp. 319-329 in *Logic, Methodology, and Philosophy of Science*, ed. E. Nagel, P. Suppes, and A. Tarski, Stanford University Press (1962).
- [7] C. A. B. Smith, "Personal probability and statistical analysis," *J. Royal Statistical Society, Series A* 128, pp. 469-499 (1965). With discussion. See also "Personal probability and statistical analysis", *J. Royal Statistical Society*,

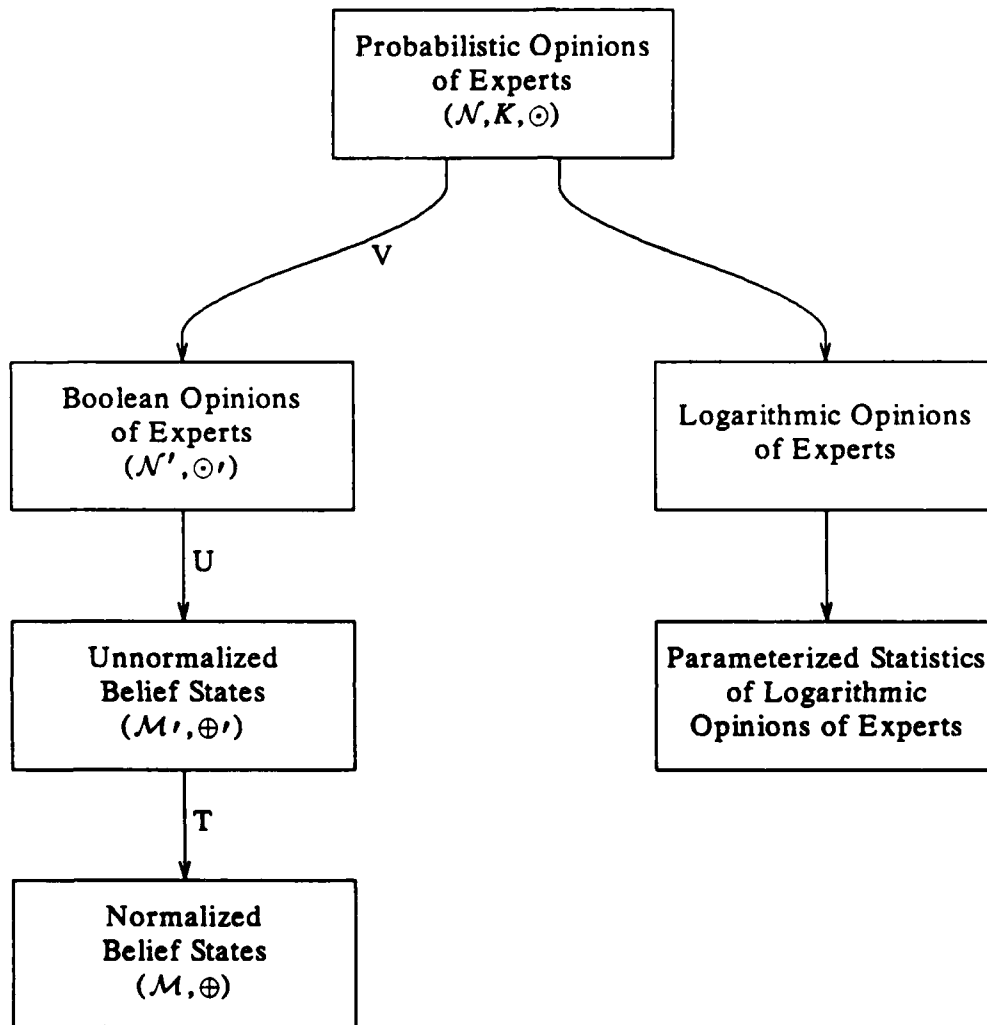


Figure 2. Names of spaces and maps between them. Each box contains the name of space, and each arrow is a homomorphism that maps onto the next space, thereby defining a representation. Note that the left branch gives the spaces involved in the interpretation of the Dempster/Shافر theory of evidence, whereas the right branch is the alternative method for combining evidence presented in Section 5.

Series B 23, p. 1-25.

- [8] A. P. Dempster, "A generalization of Bayesian inference," *Journal of the Royal Statistical Society, Series B* 30, pp. 205-247 (1968).
- [9] G. Shafer, "Allocations of Probability," Ph.D. dissertation, Princeton University (1973). Available from University Microfilms, Ann Arbor, Michigan.

A Statistical Viewpoint on the Theory of Evidence

- [10] G. Shafer, "A theory of statistical evidence," in *Foundations and Philosophy of Statistical Theories in the Physical Sciences, Vol II*, ed. W.L. Harper and C.A. Hooker, Reidel (1975).
- [11] G. Shafer, *A mathematical theory of evidence*, Princeton University Press, Princeton, N.J. (1976).
- [12] D. V. Lindley, "Scoring rules and the inevitability of probability," *Int. Stat. Rev.* **50**, pp. 1-26 (1982).
- [13] P. M. Williams, "On a new theory of epistemic probability," *British Journal for Philosophy of Science* **29**, pp. 375-387 (1978).
- [14] G. Shafer, "Constructive probability," *Synthese* **48**, pp. 1-60 (1981).
- [15] G. Shafer, "Lindley's paradox," *Journal of the American Statistical Association* **77**, pp. 325-351 (1982). (Includes commentaries).
- [16] D. H. Krantz and J. Miyamoto, "Priors and likelihood ratios as evidence," *Journal of the American Statistical Association* **78**, pp. 418-423 (1983).
- [17] T. D. Garvey, J. D. Lowrance, and M. A. Fischler, "An inference technique for integrating knowledge from disparate sources," *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, pp. 319-325 (1981).
- [18] J. Gordon and E. H. Shortliffe, "The Dempster-Shafer theory of evidence," in *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, ed. B. G. Buchanan and E. H. Shortliffe, Addison-Wesley, Reading, Massachusetts (1984).
- [19] J.A. Barnett, "Computational methods for a mathematical theory of evidence," *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, pp. 868-875 (1981).
- [20] O. D. Faugeras, "Relaxation labeling and evidence gathering," *Proceedings of the 6th International Conference on Pattern Recognition*, pp. 405-412 IEEE Computer Society, (October 19-22, 1982).
- [21] L. Friedman, "Extended plausible inference," *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, pp. 487-495 (1981).
- [22] J. Gordon and E. H. Shortliffe, "A method of managing evidential reasoning in a hierarchical hypothesis space," Stanford Computer Science Department Technical Report (1984).
- [23] T. M. Strat, "Continuous belief functions for evidential reasoning," *Proceedings of the National Conference on Artificial Intelligence*, pp. 308-313 (1984).
- [24] Michael S. Landy and Robert A. Hummel, "A brief survey of knowledge aggregation methods," NYU Robotics Report 51 (September, 1985). Submitted to the International Conference on Pattern Recognition, to be held October, 1986.
- [25] Robert A. Hummel and Steven W. Zucker, "On the foundations of relaxation labeling processes," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-5**, pp. 267-287 (May, 1983).
- [26] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and*

Machine Intelligence 6, pp. 721-741 (1984).

- [27] J. A. Anderson, "Distinctive features, categorical perception, and probability learning: Some applications of a neural model," *Psychological Review* 84, pp. 413-451 (1977).
- [28] George Reynolds, Len Wesley, Deborah Strahman, and Nancy Lehrer, "Converting feature values to evidence," Computer and Info Science Dept, University of Massachusetts at Amherst Technical Report (November, 1985). In preparation.
- [29] Eugene Charniak, "The Bayesian basis of common sense medical diagnosis," *Proceedings of the AAAI*, pp. 70-73 (1983).

END

5-87

DTIC